DOCUMENT RESUME

ED 443 840 TM 031 485

AUTHOR Hohn, Alisabeth M.; Veitch, William R.
TITLE How Much Is Too Much? Multiple Measures in Practice.
PUB DATE 2000-04-27
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Opinion Papers (120) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Accountability; Elementary School Students; *Evaluation Methods; Primary Education; Program Evaluation; Research Utilization; State Programs; *Testing Programs
IDENTIFIERS *Colorado Springs Public Schools CO; *Multiple Measures Approach

ABSTRACT

This paper describes the experiences of Colorado Springs School District Eleven in using multiple measures to determine if third graders were reading at "grade level." The use of multiple measures may be threatened by the rapidly increasing pressures about assessment and accountability in Colorado that teachers are beginning to feel that their instruction time is being replaced with testing. The increase in state-mandated testing and accountability based solely on state test scores puts added pressure on school districts with established testing programs to systematically decrease district testing programs. Multiple measures thinking requires some redundancy in the service of better decision making, but at some point redundancy becomes overkill. Two areas seem important to this discussion. One is providing accurate information to consumers and the other is training consumers to interpret data effectively. The question of how much testing is too much can only be answered when each test and the information it provides are clearly understood. If a test does not meet criteria of validity, fairness, credibility, and utility, it should be reconsidered. Appendixes contain a description of the Colorado Student Assessment Program, a chart to reflect what assessment audiences want to know, and a table of time spent on large-scale testing. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 443 840

How Much is too Much?: Multiple Measures in Practice

A paper in a symposium on Multiple Validations of Student Performance
Annual Meeting of the American Educational Research Association
New Orleans, LA
April 27, 2000

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

By

Alisabeth M. Hohn, Ph.D. and William R. Veitch, Ph.D.
Colorado Springs School District Eleven

A. M. Hohn

1

In a symposium on multiple measures last year (Hohn and Veitch, 1999), we discussed our school district's response to a state requirement of using multiple measures to determine if our third graders are "reading at grade level" as part of the Colorado Basic Literacy Act. It is not a particularly "high stakes" decision for individual students, in that retention or additional instruction requirements such as summer school are not presently attached to the "grade level" determination. Instead, students below grade level are given an Individual Literacy Plan (ILP), and must have their literacy level monitored every six months until they catch up to their grade-level peers or exit the system. In last year's paper, we discussed which measures we chose and how we chose them, how cut scores were determined on each instrument, and how we arrived at our combining rules to determine "grade level."

After conducting several empirical studies and consulting regularly with teachers and other data consumers, we arrived at a decision rule that students need to be proficient on two of three required indicators in order to be considered "reading at grade level or above." The decision rule made sense and was easily explainable to teachers and parents (e.g., a student could have a bad day for one of the tests and still "pass."). The process and outcome aligns with CRESST's recommendation that accountability systems should be guided by validity, fairness, credibility, and utility (Baker quoted in Lewis, 2000).

In the year since these decisions were made, teachers and principals in the district have largely embraced the idea that decisions about individual students should be made after evaluating several pieces of evidence. In fact, our teachers felt so strongly that multiple indicators are important, they decided to require evidence from multiple sources to remove a student from an ILP once he or she has been identified. The law and associated rules do not specify criteria for removal from an ILP, but our teachers and principals chose rigorous and specific criteria anyway.

This is an enlightened perspective, and clearly is in the best interest of students, but "times, they are a changin'" in Colorado, and sadly, our enthusiasm and opportunity may be short-lived. High stakes accountability for schools and districts is forthcoming and state-wide

TM031485

BEST COPY AVAILABLE

standards tests are taking on an importance we have never seen in Colorado before, but we are not alone. According to a recent article in Education Week (Olson, April 5, 2000):

- 49 states have adopted standards in at least some academic areas
- 48 states have testing programs designed to measure how well students perform on standards
- 21 states plan to issue overall ratings of their schools based largely on their students' performance.
- At least 18 states have the authority to close, take over, or overhaul schools that are identified as failing.

High stakes for students are also becoming more common:

- In 20 states, students must pass a test to earn a diploma. In the next three years the number will increase to 28.
- At least half a dozen states plan to tie student promotion to test results.
- US Senator Paul Wellstone plans to introduce legislation that would require states and districts to use multiple measures of performance if they are going to use standardized tests to make high-stakes decisions about students, such as graduation or promotion.

Pressures about assessment and accountability are growing rapidly in Colorado because of a new accreditation law and an education reform package signed into law last week. The education reform package expands the state testing program (called CSAP) from 8 tests administered in 2000, to 19 tests in 2001, to 27 tests in 2002 (see Appendix A for details of grade levels and content areas to be tested). The law also includes school report cards beginning next year with normative-based letter grades for schools calculated *exclusively* from CSAP scores. Banks (1994) accurately described the conflicting purposes of politicians and educators now being felt in Colorado in noting that, "In the political arena, information about student performance leads to demands for change, and policymakers use humiliation, competition, and sanctions to inspire those who fall behind. In the educational arena, the importance of test data depends upon the extent to which it helps teachers improve their instruction."

Our teachers feel that their instruction time is being replaced with testing. They fear that test scores will be used to evaluate their performance, and feel pressure to get their students to perform to high standards quickly in an environment in which the governor and legislature believe that poverty is simply an excuse for poor performance, and are unwilling to allocate state funds to lessen the problems associated with socioeconomic differences in achievement. In some schools, teachers and principals appear energized by the challenge. They want to use all data available to them to help improve student achievement. In other schools, learned helplessness has already set in.

The increase in state-mandated testing and accountability solely based on state test scores puts increased pressure on school districts with established testing programs such as ours, to systematically decrease district testing requirements. A quick response to this pressure would be short-sighted, and would likely backfire, so we find ourselves in a difficult

position. The pressure to react is considerable; people want relief immediately. However, a reduction in district testing could dramatically reduce our ability to provide schools with multiple indicators of achievement and place undue confidence in and reliance upon state test scores. Also, our teachers are just learning how to interpret data from multiple sources to make good instructional decisions and how to use large-scale data to accurately evaluate school effectiveness over time. *Multiple measures thinking requires some redundancy in the service of better decision-making, but at what point does redundancy become overkill?*

What is an Assessment Unit to do in this situation? The governor made it clear that large-scale tests and very public accountability in education are non-negotiable, so we had to decide what parts of the system we can influence. Preparing this paper has afforded us an opportunity to reflect on the concerns we hear from the field on a regular basis. We have come to the conclusion that many concerns of teachers, parents, and principals are based on a lack of accurate information about district and state tests, an egocentric view of purposes and audiences for assessment results, and a lack of skill in interpreting results and making instructional decisions based on them. If the experience of teachers with the TAAS test Texas is any indication (Gordon and Reece, 1997), the problem is not going to go away on its own and even more instruction time will be lost to test preparation divorced from good instructional practice.

We know that multiple measures help teachers make better decisions for kids, but we are in danger of losing our ability to do so. Already, teachers and principals are referring to themselves as "D" schools or "C" schools, although the report cards won't come out for another year (thanks to "unofficial" -and inaccurate- letter grades calculated by our local newspaper). If we wish to maintain a large-scale assessment system that includes more than state mandated testing, we have a lot of work to do. Unless we can empower data consumers with knowledge and skills they do not yet possess and demonstrate the predictive validity of our district-wide assessments, the district testing program will likely disappear. Our work on this topic is focused in two areas:

Providing Accurate Information to Consumers. What are the purposes and who are the audiences for the various assessment results? How much time is spent on testing now? What are the possible scenarios for the future? What data already exist in the district? How stable have state testing programs been in the past? What are possible consequences of a scaled-back assessment system?

Training Consumers how to Interpret Data Effectively. How can the data be used instructionally? How well do district assessments predict performance on state assessments? If certain assessments are eliminated, will information gaps be created? How do multiple measures tell a different story than single indicators of achievement?

In an era of expanding accountability mandates, we have found it important to do a lot of listening and a lot of training. Our teachers feel a loss of instructional autonomy and

pressure to do more in less time. Tests are often blamed for this change for several reasons, and assessment staff can be a convenient target for their anger. Our teachers are not used to the idea that other audiences might be interested in monitoring student achievement, so we have developed a role-playing exercise to raise awareness (see Appendix B). Together, we generate a list of audiences for assessment results and identify what each audience wants to know about student achievement. This activity allows us to highlight the differences between large-scale and classroom data and the strengths and weaknesses of each. Then, we examine each instrument in the testing program and identify primary and secondary audiences for the results. How does each instrument help answer the questions from each audience? From this, we are able to point out how no test can answer all the questions of each audience equally well and that there are tradeoffs inherent in any test (e.g., test length and detail of reports).

A useful follow-up to this activity is a discussion of the actual amount of time students are engaged in tests for various purposes. If you were to ask teachers to estimate the amount of instruction time dedicated to large-scale assessment, what would they say? Phelps (1996) notes that, "a blanket assertion that U.S. students are 'the most heavily tested on earth' has some validity problems." His data suggest that "on some types of tests, not only were U.S. students not the most heavily tested on earth, they were *the least heavily tested*" among the nations in his study. In response to the assertion that there is too much testing in our district, we created a document to show exactly how much time is spent on large-scale tests for each grade level, and how much there will be when the new tests required by the state are added (see Appendix C). The percentage of time spent on large-scale tests is surprisingly small (approximately 2% of instruction time), even with the added requirements.

This document has been most positively received when it is shared *after* having the discussion about purposes and audiences for assessment results. Still, we have had some surprisingly angry responses from some teachers. So, before you construct your own matrix, be prepared; we have discovered that to teachers "testing time" clearly means much more than the actual time students are taking the tests. Their concerns fall into several categories:

- The calculations include only test administration time, not test preparation time.
- Any day that includes a large-scale test may be considered an entire day of instruction lost. A three-class period test is expressed as a week of instructional time lost.
- Individually administered instruments are shown with the amount of time required of the student, not of the teacher (e.g., 30 minutes vs. 30 minutes x 24 students).
- Some have argued that the contact hours used to determine percentages of testing time (based on contact hours required by law) are an overestimate of the actual time students are engaged in instruction.

Roughly two percent of instruction time for assessment with a primary audience other than the classroom teacher is minimal compared to the time teachers spend assessing students to calculate grades. Even though the primary audience for results is someone other than the classroom teacher, many large-scale assessments contain information that can inform

instruction, especially if the assessments are aligned with standards, if the tests are sensitive to differences in instructional practice, and if the results are returned in a timely manner (Baker and Linn, 2000). If the teacher is not skilled in using the results in conjunction with their classroom data, they may be more inclined to resent giving some class time to it. Two examples are described below:

- In the elementary grades, many teachers feel that valuable instruction time is being lost because the Literacy Act requires the use of individually administered assessments (especially in the fall when students have not yet learned the routines of the classroom and are not very independent workers). Some of our teachers, who are skilled at differentiating instruction, have noted that they would not know what to teach without doing this kind of assessment at the beginning of the year, but many others do not see value in the data they are collecting.

- At the high school level, the only assessment the district has administered systematically in many years is the Riverside's Tests of Academic Proficiency (TAP). It is not instructionally relevant or sensitive to programmatic variation. Beginning next year, 9th, 10th, and 11th graders will take state-wide tests, so our high schools are trying to figure out how they will manage the logistics of test administration. Also, because they have not received instructionally relevant results in the past, the new tests are perceived as a waste of time. Finally, because the tests will have very low stakes for students (not tied to graduation), teachers and administrators are concerned about student motivation, and therefore do not expect the results to be a valid summary of achievement in their school.

In this environment, the idea that more testing is better is laughable. However, we are finding that once our teachers feel that their concerns have been heard, many are surprisingly willing to learn how they might use data more effectively. To maximize the usefulness of data, teachers need to become skilled at differentiating instruction. They need to develop skills in: how to document and prepare for students with different instructional needs, how to interpret assessment results from multiple sources, and what instructional strategies will help students with various identified needs. Administrators need to be instructional leaders who understand what different tests measure and what they can and cannot tell us.

According to Banks (1994), "The purpose of local assessments – to improve instruction – may differ from the purpose of externally mandated assessments." This is clearly the case in Colorado. However, assessment unit staff can help data consumers make appropriate and instructionally relevant inferences from large-scale data and guide teachers toward developing appropriate test preparation practices that minimally detract from regular instruction (recommendation in Gordon and Reese, 1997).

In a survey study of over 100 Texas teachers, Gordon and Reese (1997) found that, "High-stakes testing has become the object rather than the measure of teaching and learning, negatively affecting curriculum, teacher decision making, instruction, student learning, school climate, and student motivation." In the introduction to this paper, we

shared a story of how a multiple measures approach to decision-making is *positively* affecting teaching and learning in our district, but high stakes testing is on the way and threatens the kind of thoughtful considerations our teachers are beginning to develop about data.

Earlier in this paper, we made a statement about multiple measures and then posed an important question; we stated that multiple measures thinking requires some redundancy, but at what point does redundancy become overkill? We believe that the question can be answered only after one clearly understands each test and what information each provides to various audiences. Once one understands how to interpret the data appropriately from different perspectives, it has been our experience that the "too much testing" argument tends to go away. And when the stakes for students are high, using multiple measures becomes increasingly desirable.

So how much is too much? We would argue that when a test does not meet criteria of validity, fairness, credibility, and utility, it should be reconsidered. In a multiple measures environment, we would add that each test should provide corroborative evidence of the other indicators along with some unique information. Ideally, a district's assessment system will be both efficient and comprehensive.

REFERENCES

Baker, E. L., & Linn, R. L. (2000, Winter). Alignment: Policy goals, policy strategies, and policy outcomes. CRESST Line, 1-3.

Banks, K. E. (1994). Assessment's conflicting purposes, conflicting politics: impact on local school systems. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Glaser, R. and E. Silver. (1994). Assessment, testing, and instruction: retrospect and prospect. CSE Technical Report 379, UCLA: CRESST.

Gordon, S. P., and Reese, M. (1997). High-stakes testing: Worth the price? Journal of School Leadership. 7(4), 345-368.

Grasmick, N. S. (1996). The politics of large scale assessment. Paper presented at the Conference on Large Scale Assessment, Colorado Springs, CO.

Hohn, A. M., and Veitch, W. R. (1999). The Colorado Basic Literacy Act. Paper presented at the American Educational Research Association, Montreal, Canada.

Lewis, A. (2000, Winter). 1999 CRESST Annual Conference benchmarks for accountability: Are we there yet? CRESST Line. 3-5.

Olson, L. (2000, April 5). Worries of a standards 'backlash' grow. Education Week, 19 (30), 1, 12-13.

Phelps, R. P. (1996). Are U.S. students the most heavily tested on earth? Educational Measurement: Issues and Practices, 15(2), 19-27.

Appendix A

Colorado Student Assessment Program (CSAP)
Expansion of the Testing Program from 2000 - 2002

	Year		
Grade	**2000**	**2001**	**2002**
K			
1			
2			
3	R	R	R,W
4	R,W	R,W	R,W
5	M	R,M	R,W,M
6		R	R,W,M
7	R,W	R,W	R,W,M
8	M,S	R,M,S	R,W,M,S
9		R	R,W,M
10		R,W,M	R,W,M
11*		R,W,M,S	R,W,M,S
12			
Total	**8**	**19**	**27**

R = Reading, W = Writing, M = Math, S = Science

*Note: All 11th graders will take the ACT test

Features of the CSAP Tests:

- Measure Colorado Standards
- Each test is approxomately 3 hours in length
- Tests are mixed format (multiple choice and constructed response)
- Turn around time for results is slow, but is expected to improve
- Students receive a scale score and proficiency level on each test
- Tests will be vertically equated beginning next year
- Tests were not designed to be high stakes
- Some sample items are released after each administration for future test prep

Appendix B

Audiences for Assessment Results
What do they want to know?

Audience	What do they want to know?
Public	
School Board	
Central Administration	
Dept. of Instruction	
Principal	
Teacher	
Parent	
Student	

How much time is spent on testing?
1999-2000 School Year

Colorado Springs School District Eleven

Hours of Actual Student Testing Time*

Assessments	Grade Level												
	K	1	2	3	4	5	6	7	8	9	10	11	12
CSAP (State Assessment)				2	6	3		6	6				
DALT (District Ach. Levels Tests)				6	6	6	6	6	6				
K-3 Literacy Assessments	1	1.5	1	1	1a								
ITBS Survey Battery						1.5		1.5					
TAP Survey Battery											1.5		
District Writing Assessment				2			2			2			
End of Year Reading Test										1			
Total hours of testing time													
Minimum	1	1.5	1	11	12	10.5	8	13.5	12	3	1.5	0	0
Maximum	1	1.5	1	11	13	10.5	8	13.5	12	3	1.5	0	0
Total Contact Hours in School													
Minimum	452	720	720	720	720	720	720	720	720	720	720	720	720
Maximum	478	990	990	990	990	990	1080	1080	1080	1080	1080	1080	1080
% of Testing Time													
Minimum	0.2%	0.2%	0.1%	1.1%	1.2%	1.1%	0.7%	1.3%	1.1%	0.3%	0.1%	0.0%	0.0%
Maximum	0.2%	0.2%	0.1%	1.5%	1.8%	1.5%	1.1%	1.9%	1.7%	0.4%	0.2%	0.0%	0.0%

* Actual student testing time does not include materials preparation or test prep training.

Notes:

a: A formal literacy assessment is administered to students who are reading below grade level in grade 4.

How much will be spent on large-scale testing: A scenario
2001-2002 School Year

Colorado Springs School District Eleven

Hours of Actual Student Testing Time*

Assessments	Grade Level												
	K	1	2	3	4	5	6	7	8	9	10	11	12
CSAP (State Assessment)				5	6	9	9	9	12	9	9	5a	
DALT (District Ach. Levels Tests)				3	3	3	3	3	3				
K-3 Literacy Assessments	1	1.5	1	1	1b	1b	1b						
Total hours of testing time													
Minimum	1	1.5	1	9	9	12	12	12	15	9	9	5	0
Maximum	1	1.5	1	9	10	13	13	13.5	15	9	9	5	0
Total Contact Hours in School													
Minimum	452	720	720	720	720	720	720	720	720	720	720	720	720
Maximum	478	990	990	990	990	990	1080	1080	1080	1080	1080	1080	1080
% of Testing Time													
Minimum	0.2%	0.2%	0.1%	0.9%	0.9%	1.2%	1.1%	1.1%	1.4%	0.8%	0.8%	0.5%	0.0%
Maximum	0.2%	0.2%	0.1%	1.3%	1.4%	1.8%	1.8%	1.9%	2.1%	1.3%	1.3%	0.7%	0.0%

* Actual student testing time does not include materials preparation or test prep training.

Notes:

a. The ACT will be administered to all 11th graders. The duration of the test is unknown and was estimated at 5 hours.

b. A formal literacy assessment is administered to students who are reading below grade level in grade 4, 5, and 6.